MATHEMATICS OF DATA SCIENCE

November 21, 2023

Today: More on concentration inequalities

⚠ Next lecture: Monday November 27
5.15pm - 6.45pm

Previously on MDS

↳ Random variables (continuous for sake of generality)

Y random variable is defined by a density (aka a probability distribution)

$$p : \mathcal{Y} \subseteq \mathbb{R} \longrightarrow \mathbb{R}_+$$
$$y \mapsto p(y) \geq 0$$

"support" of the distribution (and thus of Y)

$$\int_{\mathcal{Y}} p(y)\,dy = 1$$

↳ Markov's inequality

If Y is a random variable such that $\mathbb{E}[|Y|] < \infty$, then $\forall \varepsilon > 0$, $\mathbb{P}(|Y| \geq \varepsilon) \leq \frac{\mathbb{E}[|Y|]}{\varepsilon}$

Chebyshev's inequality

If Y is a random variable such that $\mathrm{var}[Y] \left(= \mathbb{E}[(Y - \mathbb{E}[Y])^2]\right) < \infty$, then

$$\forall \varepsilon > 0, \quad \mathbb{P}(|Y - \mathbb{E}[Y]| \geq \varepsilon) \leq \frac{\mathrm{var}[Y]}{\varepsilon^2}$$

$Y - \mathbb{E}[Y]$: centered random variable

$\mathbb{E}[Y - \mathbb{E}[Y]] = 0$

(1) iid variables

↳ Two random variables Y_1 and Y_2 are called identically distributed if they have the same probability distribution p (and thus the same support)

↳ For any pair of random variables Y and Z with supports $\mathcal{Y}$ and $\mathcal{Z}$, respectively, the product (Y, Z) is a random variable and it is defined by a product distribution

$$p : \mathcal{Y} \times \mathcal{Z} \longrightarrow \mathbb{R}_+ \quad \text{such that}$$
$$(y, z) \mapsto p(y, z) \qquad \int_{\mathcal{Y}} \int_{\mathcal{Z}} p(y, z)\, dy\, dz = 1$$

- the marginal distribution of Y corresponding to p is

$$p_Y : \mathcal{Y} \longrightarrow \mathbb{R}_+$$
$$y \mapsto p_Y(y) = \int_{\mathcal{Z}} p(y, z)\, dz$$

$\Rightarrow$ Expected value $\mathbb{E}_Y[\cdot]$ and $\mathrm{Var}_Y[\cdot]$ are defined according to p_Y

- Similarly, the marginal distribution of Z is

$$p_Z : \mathcal{Z} \to \mathbb{R}_+$$
$$z \mapsto p_Z(z) = \int_{\mathcal{Y}} p(y, z)\, dy$$

Def: Two random variables Y and Z are **independent** if

$$\forall (y, z) \in \mathcal{Y} \times \mathcal{Z}, \quad p(y, z) = p_Y(y) \times p_Z(z)$$

(p: joint distribution (aka product distribution); p_Y, p_Z: marginal distributions)

↳ If Y and Z are independent random variables with the same probability distribution, we say that they are **iid** (independent, identically distributed)

⇒ Many results in statistics consider iid samples and derive concentration inequalities

(2) Hoeffding inequalities

↳ Hoeffding's inequality applies originally to Bernoulli variables, that take values 0 or 1

Bernoulli variable: Random variable Y, $Y = \{0, 1\}$

$\mathbb{P}(Y=0) = 1-p$ $\quad \mathbb{P}(Y=1) = p$

$p \in [0,1]$

Symmetric Bernoulli variable / Rademacher variable

Random variable Y, $Y = \{-1, 1\}$

$$\mathbb{P}(Y=-1) = \mathbb{P}(Y=1) = \frac{1}{2}$$

Theorem (Hoeffding's inequality)

Suppose that $Y_1, \dots, Y_N$ are iid Rademacher variables (with $N \geq 1$). Let $a \in \mathbb{R}^N$. Then, for any $\varepsilon \geq 0$,

$$\mathbb{P}\left(\sum_{i=1}^{N} a_i Y_i \geq \varepsilon \right) \leq \exp\left(\frac{-\varepsilon^2}{2\|a\|^2} \right) \in [0,1] \text{ for } \varepsilon \geq 0$$

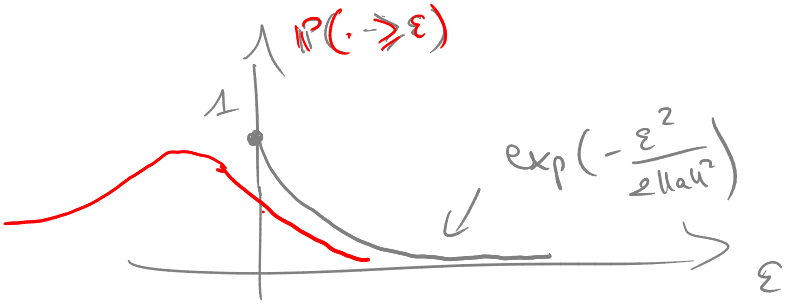
- $\mathbb{P}\left(\sum_{i=1}^{N} a_i Y_i \geq \varepsilon \right)$: probability that the linear combination exceeds ε
- $\sum_{i=1}^{N} a_i Y_i$: linear combination of the random variables
- $\exp\left(\frac{-\varepsilon^2}{2\|a\|^2} \right)$: exponential function $\to 0$ as $\varepsilon \to \infty$; fast decrease as ε increases

→ Hoeffding's inequality shows that

$$\mathbb{P}\left(\sum_{i=1}^{N} a_i Y_i \geq \varepsilon\right) \ll \mathbb{P}\left(\sum_{i=1}^{N} a_i Y_i < \varepsilon\right)$$

for sufficiently large values of ε

⇒ We say that the distribution of $\sum_{i=1}^{N} a_i Y_i$ is light-tailed or that it concentrates (around its mean)



Remarks

- The proof of Hoeffding's uses Markov inequality applied to $\mathbb{P}\left(\exp\left(\lambda \sum_{i=1}^{N} a_i Y_i\right) \geq \exp(\lambda t)\right)$ for any $\lambda > 0$ to obtain a bound as a function of λ, then it optimizes over λ to find the best bound (convex optimization)

- If $a = \begin{bmatrix} 1/N \\ \vdots \\ 1/N \end{bmatrix} \in \mathbb{R}^N$, then $\|a\| = \sqrt{\sum_{i=1}^{N} \left(\frac{1}{N}\right)^2} = \sqrt{\frac{N}{N^2}} = \frac{1}{\sqrt{N}}$

and Hoeffding's inequality becomes

$$\mathbb{P}\left(\frac{1}{N}\sum_{i=1}^{N} Y_i \geq \varepsilon\right) \leq \exp\left(\frac{-N\varepsilon^2}{2}\right)$$

$\frac{1}{N}\sum_{i=1}^{N} Y_i$: empirical estimate of $\mathbb{E}[Y_1] = \dots = \mathbb{E}[Y_N]$

$\exp\left(\frac{-N\varepsilon^2}{2}\right) \to 0$ as $\varepsilon \to \infty$ or as $N \to \infty$

For small ε, increasing N leads to a better empirical estimate of the mean

- Hoeffding's is non-asymptotic because it is valid for any N
- Hoeffding also works for other types of random distributions.

If $Y_1, \dots, Y_N$ are iid variables such that $m \le Y_i \le M$ for some deterministic values $m \le M$ then

$$\mathbb{P}\left(\sum_{i=1}^{N} (Y_i - \mathbb{E}[Y_i]) \ge \varepsilon \right) \le \exp\left(-\frac{2\varepsilon^2}{N(M-m)} \right)$$

$\Rightarrow$ this is also called Hoeffding's inequality

Q: What kind of random variables have a similar "concentration property" / "light-tail property"?

(3) Sub-gaussian (random) variables

Def: A random variable Y is called sub-gaussian if it exists $K > 0$ such that $\mathbb{E}\left[\exp\left(\frac{Y^2}{K^2}\right) \right] \le 2$,

which is equivalent to (light tailed distribution)

$$\mathbb{P}(|Y| \ge t) \le 2\exp\left(-\frac{ct^2}{K^2} \right) \quad \text{for some } c > 0.$$

The smallest constant K satisfying these properties defines a norm on random variables, and is denoted by $\|Y\|_{\psi_2}$

Examples
- $Y \sim N(0, \sigma^2)$ (Gaussian variable) is sub-gaussian with $\|Y\|_{\psi_2} \le K_1 \sigma$, $K_1 = \|Y_0\|_{\psi_2}$ with $Y_0 \sim N(0,1)$

- Symmetric Bernoulli is sub-gaussian with $\|Y\|_{\psi_2} = \frac{1}{\sqrt{\ln 2}}$

- Bounded random variable is sub-gaussian with
$$\|Y\|_{\psi_2} \leq \frac{1}{\sqrt{\ln 2}} \|Y\|_\infty$$
$\|Y\|_\infty$: $\max\{|y| : y \in Y\}$

- Others (...)

Theorem: (Hoeffding's inequality)

Let $Y_1, \dots, Y_N$ be N independent, sub-gaussian random variables with zero mean ($\mathbb{E}[Y_1]=0, \dots, \mathbb{E}[Y_N]=0$)

Then, $\forall \varepsilon \geq 0$,

$$P\left(\left| \sum_{i=1}^N Y_i \right| \geq \varepsilon \right) \leq 2 \exp\left(\frac{-c\varepsilon^2}{\sum_{i=1}^N \|Y_i\|_{\psi_2}^2} \right)$$

$\sum_{i=1}^N Y_i$: sum of independent but not necessarily iid variables

where $c>0$ is a "universal" constant that does not depend on $Y_1, \dots, Y_N$

↳ Variations of this inequality:

- Consider $\sum_{i=1}^N a_i Y_i$ where $a = \begin{bmatrix} a_1 \\ \vdots \\ a_N \end{bmatrix} \in \mathbb{R}^N$

$$P\left(\left| \sum_{i=1}^N a_i Y_i \right| \geq \varepsilon \right) \leq 2 \exp\left(\frac{-c\varepsilon^2}{\|a\|^2 \max_{1 \leq i \leq N} \|Y_i\|_{\psi_2}^2} \right)$$

- Can derive a variant for sub-gaussian variables with nonzero mean by centering; i.e. replace Y_i with $Y_i - \mathbb{E}[Y_i]$ (which is sub-gaussian if Y_i is)

Application Consider a graph of n vertices where edges are generated between all pairs of vertices with probability p (Erdös-Rényi graph)

$\mathbb{P}(\text{edge } (i,j) \text{ is generated}) = p \Rightarrow$ Bernoulli

At every vertex, the expected number of edges is $d = (n-1)p$

Using Hoeffding-type inequalities, can show that if $d \geq C \ln(n)$ for some $C > 0$, then with probability 0.9, all vertices have degree (number of edges that include the vertex) between $0.9d$ and $1.1d$

$\Rightarrow$ the number of edges connecting one vertex to others concentrates around d